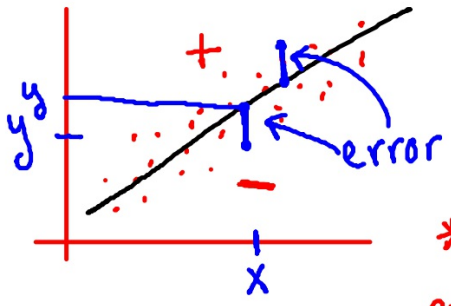
4.2 notes

Line of Best Fit



- can go thru pts. however, doesn't have to.

* All lines have errors

errors = actual y − predicted y

- **Regression Line** = **Line of best fit**

- It is a line that fits … the pattern of the data the best "average"

- The line describes how… the response variable (Y) changes as the explanatory variable (X) changes

- Used to … predict a Y-value when given an X-value

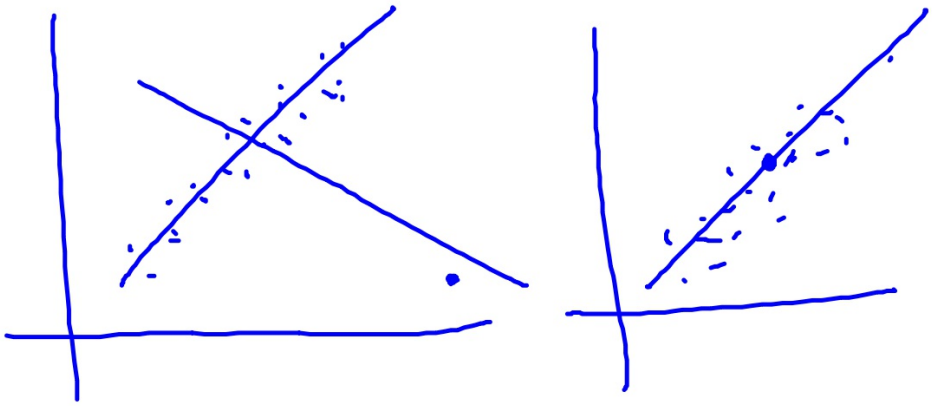
Most accurate Regression line:

- Called: Least Squares Regression Line OR LSR line OR LSRL
- Minimizes ... the sum of the squared errors
- Form: $\hat{y} = a + b(x)$ — "hat" – sample — $\hat{y} = a + bx + cx^2 + dx^3 +$

- Pieces:
  - a = intercept
  - The Y-Intercept is the … value of Y when X = 0
  - b = slope
  - The slope states that … For every change in 1 unit of the X-variable, the Y-variable changes by the slope, on average.

* Always ... goes thru the point $(\bar{x}, \bar{y})$ (averages)

* not ... resistant to outliers... the LSRL is affected by outliers



Example: A real estate agent studied the relationship between house prices and size (square footage). He found the least-squares regression line to be:

$\hat{Y} = 51912.73 + 47.734(X)$ OR $\widehat{\text{Selling Price}} = 51912.73 + 47.734(\text{Square Feet})$.

a) What is the slope? 47.734

b) Interpret the slope. For every 1 square foot of a house the selling price increases by $47.734.

c) What is the Y-Intercept? 51912.73

d) Interpret the Y-Intercept.

When square footage of a house is 0 ft², the selling price is $51912.73

LSRL on the calculator:

* STAT --> CALC --> #8: LinReg(a+bx) -->

* LinReg(a+bx) X-var, Y-var, Y1

* Y1 = VARS --> YVARS --> FUNCTION --> Y1 --> ENTER

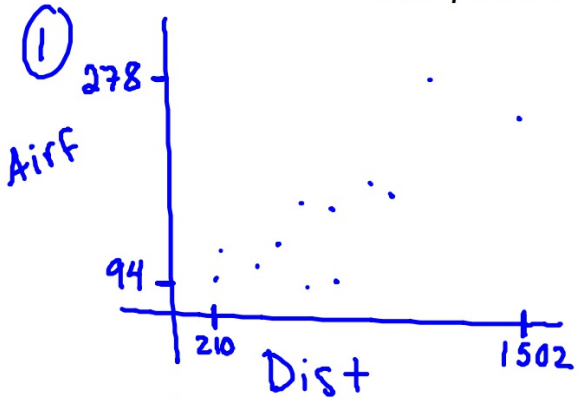
* ENTER $\hat{y} = 36.089 + 0.4474(x)$

$\rightarrow \hat{EXB2} = 36.089 + 0.4474(EXB1)$

$= 36.089 + 0.4474(78)$

70.9862%

Complete worksheet 4.2A

① 

- positive, mod. strong, linear
- no outliers

② $\widehat{\text{Airfare}} = 83.267 + 0.117(\text{Distance})$

$r = 0.795 \quad r^2 = 63.2\%$

③ airfare $= 83.267 + 0.117(370)$

$= \$126.56$

4) $\hat{y} = 0.117(1502) + 83.267$

$\hat{y} = \$259.00$

5)

370 miles	1502 miles
$e = 138 - 126.56$	$e = 258 - 259.00$
$e = \$11.44$	$e = -\$1.00$

6) $\hat{y} = 83.267 + 0.117(2842)$

$\hat{y} = \$415.78$

No, we shouldn't trust this.
The x-value is an outlier

7) $e = 198 - 415.78$

$e = -\$217.78$

8)

miles	900	901	902	903	904
airfare	188.567	188.684	188.801	188.918	189.035

9) The airfare increases $0.117 for every mile. It's the slope!

10) For every increase of 1 mile in distance the airfare increases by $0.117 on average.

11) 0.117 * 100 = $11.70

12) 63.2% of the change in the airfare is due to the change in the distance flown.

Complete worksheet 4.2B-
you will be turning this in

RESIDUALS

Ungroup RESID. You will see LIST1 (X) and LIST2 (Y).

- Create a scatterplot
- Find the LSR line, r, and r^2.
- Add the line to your plot.
- Does the line do a good job of describing the data?

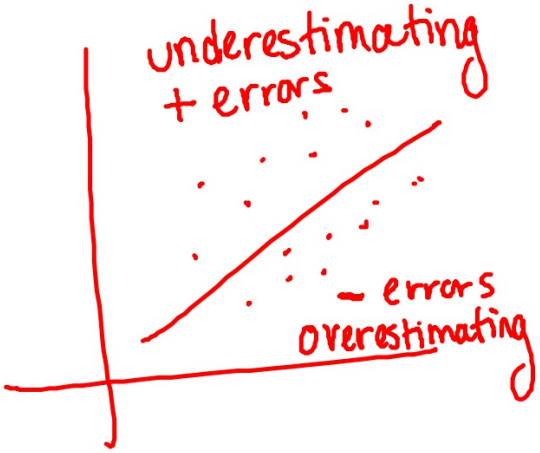
Residuals (errors):

= actual y -- predicted y ← from LSR line

= [illegible]

$$\sum residuals = 0$$

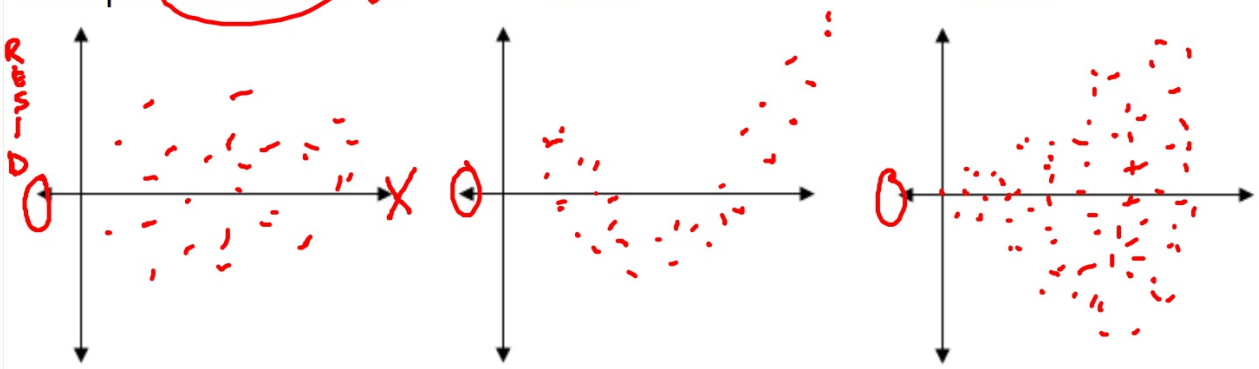
Sum



Residual Plot:

- Definition: scatterplot of X-variable vs. residuals (Y)
- Helps… to see if the linear model is appropriate for the data
- No pattern = the linear model is appropriate for the data
- Pattern = another model would be better (parabolic, cubic, exponential, etc.)

Examples: **GOOD:** scattered **BAD:** **BAD:**



Creating a residual plot on the calculator:

1- Find the line (LinReg) $\hat{y} = a + bx$

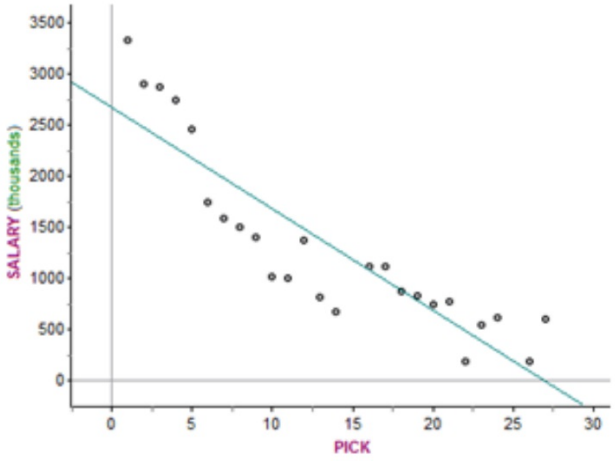
2- Scatterplot of: X-variable (X)
$_{L}$RESID (Y)

HOMEWORK: worksheet 4.2C

Complete worksheets 4.2 D, E

WORKSHEET 4.2D: ROOKIES

1)



Negative
Linear / curved
Strong (r = -0.887)
No outliers

2) $\hat{y}$ = 2,657,443.08 -- 98,957.22(X) r = -0.887 r^2 = 78.6%

$\widehat{SALARY}$ = 2,657,443.08 -- 98,957.22(PICK)

3) $r^2 = 78.6\%$

4) For every increase of 1 in the pick number from the draft there tends to be a decrease of 98,957.22 dollars in the salary.

5) When the pick (draft number) is 0, the salary is $2,657,443.08.

6) $\widehat{SALARY} = 2{,}657{,}443.08 - 98{,}957.22(12)$

$\widehat{SALARY} = \$1{,}469{,}956.44$

7) residual = actual y -- predicted y

= 1,370,000 -- 1,469,956.44 = -$99956.44

8) Overestimate. The line predicted a higher salary than the player actually received.

9) SALARY = 2,657,443.08 -- 98,957.22(15)

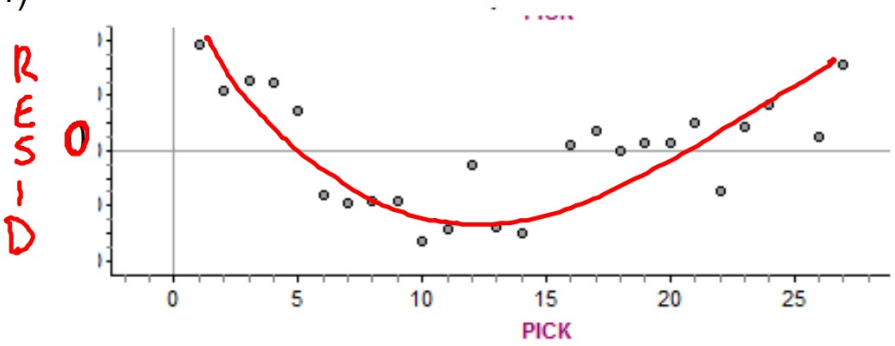
SALARY = $1,173,084.78

SALARY = 2,657,443.08 -- 98,957.22(25)

SALARY = $183,512.58

10) $98,957.22 (SLOPE)

11)



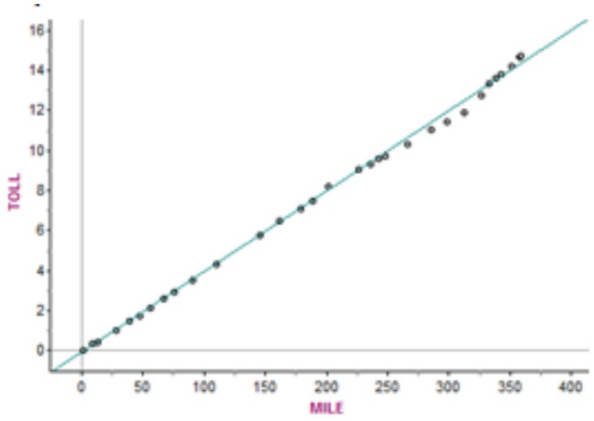
12) NO. The original plot had a slight curve, and the residual plot has a clear pattern. The correlation was high though.

13) + resid = actual > predicted UNDERestimate

14) -- resid = actual < predicted OVERestimate

WORKSHEET 4.2E

1)



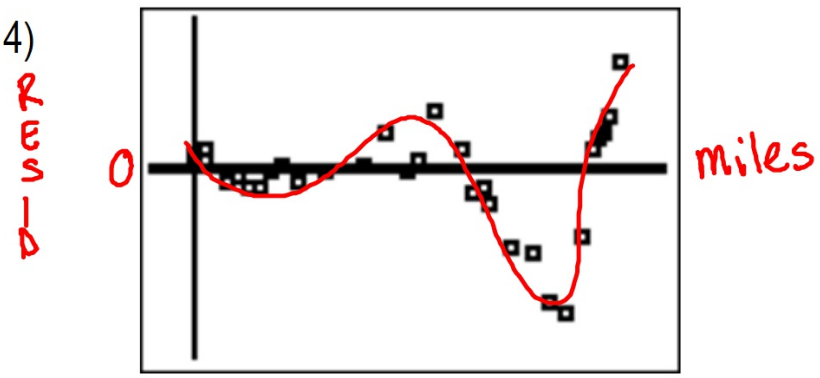
Positive
Linear
Strong (r = 0.999)
No outliers

2) $\hat{y}$ = --0.12 + 0.04(x) r = 0.999 R^2 = 99.8%

$\widehat{TOLL}$ = --0.12 + 0.04(MILE)

3) Yes, the line is a good fit. The original plot is linear, the correlation is fairly high, and the r^2 is close to 100%.

4)



5) No, the line is not the best fit for the data because the residual plot is not scattered. It has a clear pattern.

6) Yes, the line is good for the data. Just not the BEST fit for the data.

EXTRA QUESTIONS:

7) Interpret the slope in the context of the problem.

8) Interpret R^2 in the context of the problem.

9) Predict the Toll for an exit at mile 109.9.

10) The actual toll for Mile 109.9 is $4.30. Find the residual for the toll for mile 109.9.

11) Was your prediction an over or under estimate?

ANSWERS:

7) For every increase of 1 mile in the distance there tends to be an increase of $0.04 in the toll.

8) 99.8% of the change in tolls can be explained by change in the miles driven.

9) $\hat{y} = 0.04(109.9) - 0.12$

$\hat{y} = \$4.276$

10) residual = 4.30 -- 4.276 = $0.024

11) underestimate

Complete worksheet 4.2 F

Worksheet 4.2F:

1) On average in the USA, there are 1.3 people for every 1 TV in the country.

2) Not surprising. I would have expected it to be low.

3) On average in Cambodia, there are 177 people for every 1 TV in the country.

4) Not surprising. Cambodia is not a very wealthy country, so we would not expect a lot of TVs.

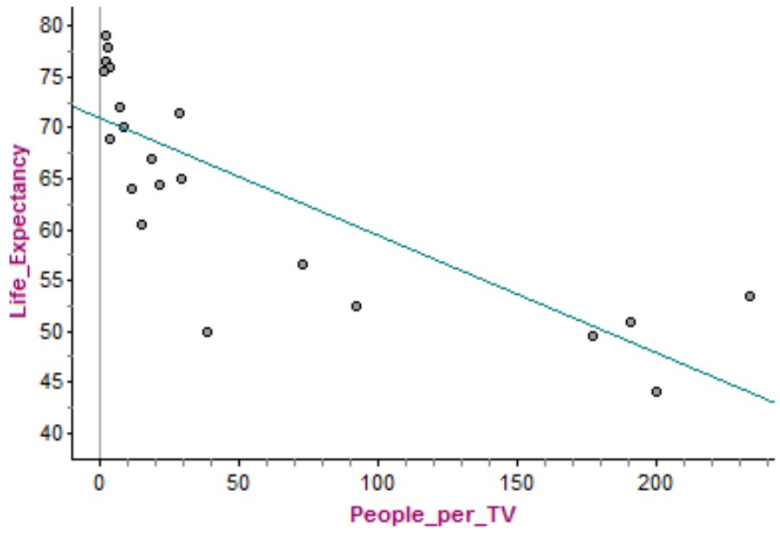
5)

LOW	HIGH
USA (1.3)	HAITI (234)
JAPAN (1.8)	ANGOLA (200)
CANADA (1.7)	UGANDA (191)

6) Why do the HIGH countries have these high numbers?

They are poorer countries. They do not have money for TVs.

7)



Negative
Curved
Moderate

8) Yes, there appears to be a curved association between the 2 variables

9) r = -0.804. The strength is moderately strong. So there is a moderately strong relationship between TV's and Life Expectancy.

10) As we increase the X variable, the Y variable will DECREASE.

11) As we decrease the X variable, the Y variable will INCREASE.

12) NOT VALID argument. More TV's do not cause people to live longer!

13) NO

14) NO. Strong correlation (association) $\neq$ causation

15) 64.6% of the change in the Life Expectancy is due to the change in the amount of people per TV's.

16) 35.4% is due to other variables.

17) MONEY!
Technology

Std. of living
Industrialization
Population

Association vs Causation:

Association = the 2 variables are related (high correlation)

Causation = one variable causes the other one to change

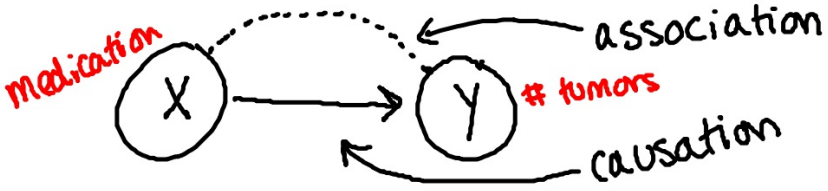
Just because there is a strong association (high correlation), does NOT mean that changing one variable causes the other one to change.

p.187

Types of Associations:

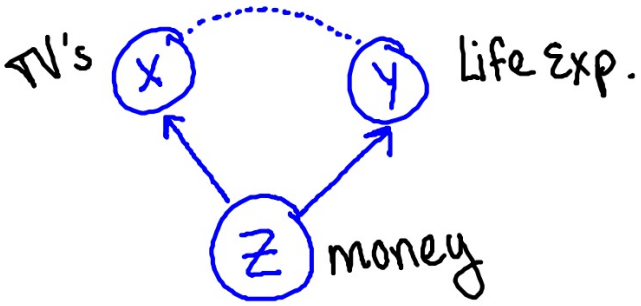
1- Causation

X causes Y to change



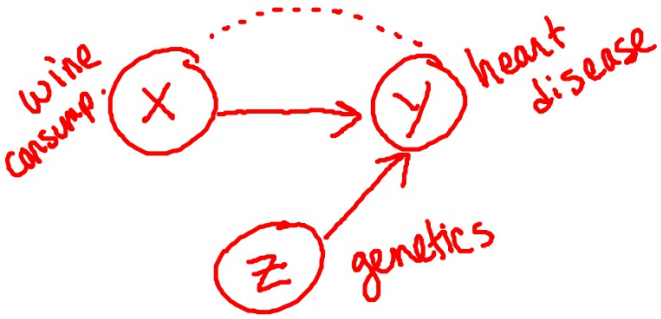
2- Common Response

the X & Y variables are both responding to a third variable

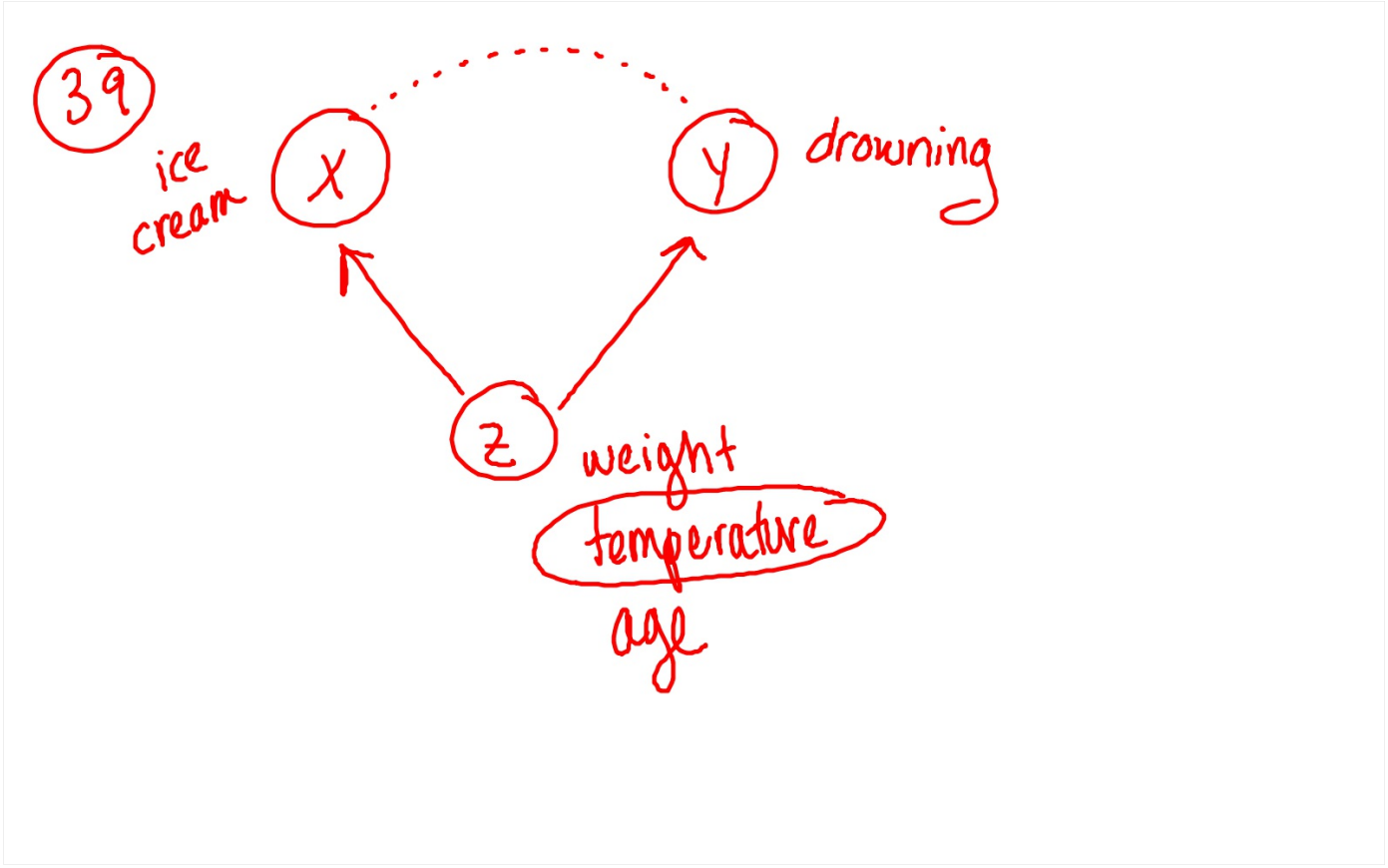


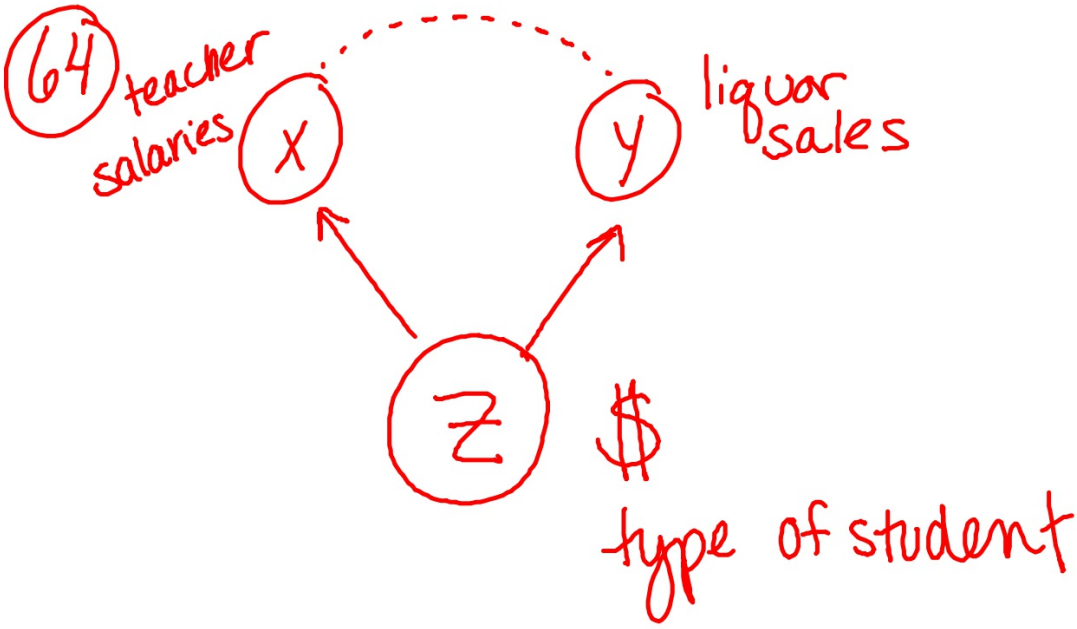
3- Confounding

the Y variable is changing because of X and a third variable



EXAMPLES:
p. 188 #39
p. 193 #53
p. 198 #64


39
ice cream
X
Y
drowning
Z
weight
temperature
age


64
teacher salaries
X
Y
liquor sales
Z
$
type of student

CW 4.2

REVIEW:

p. 197 #58, 60, 61, 63